wellcome

UKRI Economic and Social Research Council

# Realising the potential of a National Data Library

# Table of Contents

# Realising the potential of a National Data Library

The UK government has committed to the creation of a National Data Library (NDL), providing access to all public sector data and potentially other nationally-relevant information resources. Wellcome and the Economic and Social Sciences Research Council (ESRC) acknowledge the importance and opportunity of this initiative and have worked to engage the research community to understand how the NDL should be designed and run for maximum benefit. The outcome was a high degree of consensus – this paper describes the process and the key themes this expert group identified as essential to the NDL's success.

## Introduction

The UK has the potential to be a global leader in data-driven research, innovation, and public service delivery. This unique position stems from the wealth of high value data held in the UK public sector. The National Health Service (NHS) alone, used by almost all the UK's 67 million people, represents a vast source of information on individual and population health. Other sectors including education, social care, tax and benefits, and justice also hold large, unique datasets. Harnessing this type of data for research has delivered answers to important questions that have improved the lives of people here and abroad. For instance, the UK's landmark RECOVERY trial during the COVID-19 pandemic accessed around 25 different datasets and its results have saved at least a million lives globally (1, 2). Yet additional achievements are hampered by the disconnected and siloed nature of the UK's existing public sector databases. Further, researchers and policymakers face barriers when trying to access and navigate this fragmented landscape. These challenges prevent the linkage of datasets to gain important insights in health, care, and other issues affecting society. Overcoming the obstacles to this type of research will reap benefits for the public and for the UK economy, potentially by an estimated £319 billion by 2050 (3, 4).

A UK National Data Library (NDL) represents an opportunity to achieve significant societal and economic gains. It has the potential to transform scientific and health research, delivering solutions and innovations to benefit citizens in the UK and abroad, and a powerful resource that harnesses artificial intelligence (AI) to improve public services. The AI Opportunities Action Plan, published in January 2025, described the NDL as an "enormous opportunity". Yet vital steps need to be taken to realise the NDL's potential.

In early 2025, experts and stakeholders were convened by the Wellcome Trust to discuss the design and development of the NDL following its Technical White Paper Challenge on the topic. The expert group found much common ground and agreed on a number of key actions to ensure the viability of the NDL and its long-term success.

## Expert engagement

The Wellcome Trust and the Economic and Social Research Council (ESRC) launched a Technical White Paper Challenge in November, 2024, to feed into discussions about the form of the NDL. Submissions were invited from experts that proposed technical visions, architectures, and solutions for the NDL. In early January 2025, five of 21 submissions were selected by Wellcome, ERSC, and an expert evaluation committee (3, 5, 8, 9, 10). The successful papers were published on the Wellcome website on 17th January 2025. A workshop where authors of the winning submissions presented their papers to experts and other stakeholders took place at the Wellcome offices on 28th January 2025.

## The NDL's vision

Firstly, experts agreed on the need to define a clear vision for the NDL. The Labour Party Manifesto 2024 stated: "…we will create a National Data Library to bring together existing research programmes and help deliver data-driven public services, whilst maintaining strong safeguards and ensuring all of the public benefit." Although this gives some indication of the government's visions for the NDL, greater clarity is needed. For instance, meeting the needs of both researchers and operational data users may prove difficult. As Jack Hardinges and colleagues from Icebreaker One note: "The difference between data for research and data for operations is a significant one, technically, legally and commercially" (5). Defining a strong vision for the NDL is necessary not only for intended users and the public but also to inform its design. "A clear vision is a necessary design constraint" for the NDL, state Hardinges and colleagues. They recommend documenting the vision as a "clear, tightly-bound problem statement. This will help bridge the chasm between a high-level political vision and technical execution. (5)" Similarities can be drawn from the development of Open Banking in the UK, where the public were consulted about the problem they wanted addressed with their banking before the financial data innovation was designed.

## Public engagement

Past research shows UK citizens support the use of their data to benefit society, for example for medical research or service planning using health data (6). The expert group agreed that the public should be at the heart of defining the problem that the NDL will address. This approach will amplify the public support that will be essential for the success of the NDL, build on existing public trust and maintain the social contract with citizens that exists for use of their data. Policy makers can draw on the success of groups like DARE UK's Public Engagement and Data Research Initiative (PEDRI) for the NDL's public outreach (7). Several public engagements and consultations will be necessary to develop the vision for the NDL and address issues such as citizen consent as well as co-design. The expert group believes this engagement should be longitudinal and iterative to accommodate changing public views and ensure transparency for the public as development of the NDL progresses. For instance, the public may have strong views around the introduction of AI into the infrastructure that should be addressed. In engaging with the public and developing the NDL, a dialogue on data anonymisation, governance, and security is paramount for it to be considered a trustworthy enterprise by both citizens and stakeholders.

## A federated approach

The UK Government holds a wealth of data from different public sectors. Examples of high value datasets include the National Pupil Database, maintained by the Department for Education, and GP data, collected by NHS Digital. These data can prove even more valuable if they can be linked to explore relationships and interdependencies. In the past, researchers have linked such data for single analyses or in large projects where data are sent to one machine in one organisation. However, making the NDL a single, huge database has many pitfalls and is not advised by the expert group. "Putting all data on all citizens in one single database, then giving access to many users, will create unprecedented privacy risks", note Ben

Goldacre and colleagues from the Bennet Institute (8). The expert group agreed the design and delivery of the NDL should not be the responsibility of one organisation, thus reducing the possibility of infrastructure failure. Similarly, a distributed model avoids vendor lock-in, increases accountability, competition, and resilience.

A centralised approach also provides challenges for accountability and increases the risk of non-delivery. Past government data platform projects have suffered using this model. Instead, the group believes the UK Government should develop a federated, decentralised architecture for the NDL. Federation will allow the government to leverage existing data infrastructure as discussed in the next section. Different options exist for federation including a modular approach, entailing a network of single-purpose services connected together into a unified platform. This model is suggested by Goldacre and co-authors, who make two core recommendations: “Firstly: build separate, connected data centres, not one giant database. Secondly: build connected single purpose services, stitched together into a national data platform, not one monolithic organisation delivering all the work” (8). A consolidated but distributed fabric of data controllers and research-performing organisations is another federation possibility put forward by Will Browne and colleagues from Emrys Health. The controllers and organisations would work “in an open data format, connected via a common data catalogue, registry framework and citizen consent using best practice security and technical standards. The value of a data fabric approach is that it allows for true federation - central oversight, record-keeping, common tools and technology access, and local information governance and curation”, they write (9). Irrespective of the form of federation, this approach will allow the NDL to function even when one element fails or underperforms. This capability will ensure not only service continuity but would also maintain public trust.

## Utilising existing infrastructure

In developing the NDL, the government should avoid unnecessary duplication given the UK’s already crowded data ecosystem. This approach will enhance the pace, feasibility, delivery, and success of the NDL. “Intended users should be engaged early to understand the needs not met by existing data infrastructures”, recommend Hardinges and colleagues (5). Depending on its final vision for the NDL, the government can take design inspiration from existing data infrastructures such as Open Net Zero, OpenSAFELY, and UK Biobank. Wherever possible, the government should look to use, integrate, configure and connect the UK’s existing infrastructure through a federated approach. For instance, the NDL could leverage the UK’s network of Trusted Research Environments (TREs). Phase 2 of the DARE UK programme, launched in August 2024, is working on the reference implementation of a federated network of TREs and data providers. This new phase will provide the government with vital information to inform the NDL’s architecture. Already, the DARE UK programme has “conducted public and professional consultations, landscape reviews and programmes of research, development and proofs-of-concept to assemble most of the key elements needed for a federated approach to the NDL”, write M Aumgi and co-authors (3). Using and connecting existing infrastructure will also allow for significant cost savings.

## Good governance

A federated approach, however, will still require data harmonisation, governance, and common agreements. As M Amugi and colleagues at DARE UK note: “Multiple data custodians operate multiple governance regimes with multiple risk appetites. Initiating a federated project between multiple data providers requires a certain degree of harmonisation of data governance policies” (3). The expert group encouraged the government to convene sector experts to develop a consensus for each of these various standards. Of note, in the health and health-relevant data sphere, the UK Health Data Research Alliance consisting of more than 100 members from

across the UK, is working towards standards in transparency, metadata, data access and data use. This initiative should feed into the NDL's development. "Information governance standards will be a big driver of the implementation details of the federated NDL", highlight M Amugi and colleagues (3). Furthermore, the expert group felt a special purpose organisation, with completely autonomous governance, should be established to oversee accountability of the NDL's delivery and operation. This oversight would improve chances of success as well as maintain trust in the initiative from the public and stakeholders. Core governance of the NDL should be established early on and include the public, academia, industry, end users, and civil society. Good governance will reduce friction between data providers and prevent problems in the development and operation of the NDL. Browne and colleagues recommend five cross-cutting panels for oversight and governance: a technical advisory panel, a social contract panel, an information governance advisory panel, a research advisory panel, and an all-party parliamentary group. These should be "staffed with high-caliber individuals with significant domain-expertise, to provide oversight, assurance and governance" (9).

## Growth development

The complexity of the NDL should not be underestimated. Scalability and iteration across the product development cycle should be built into the NDL's design, with consideration given to communicating this feature with the public. Indeed, establishing a dedicated communications team early on will be essential for the NDL, the expert group noted. Additionally, the NDL should consist of automated, computable systems to drive efficiency. As Goldacre and colleagues state "…real opportunity lies in relentlessly automating every repeatable element, even small tasks" (8). To effectively deliver at scale, the government should also look to develop and fund orchestration. Some existing TRE infrastructures deliver fast processing and automation. For instance, the Secure e-Research Platform in Wales was able to respond to requests more quickly during the COVID-19 pandemic because it invested heavily in orchestration.

## AI readiness

To support research and innovation, the role of AI in the NDL requires special consideration. Scientists are increasingly using AI models and tools for essential tasks including discovery, cleaning, and augmenting data. Given this trend and the ambitions of the AI Opportunities Plan, the government needs to make the NDL data machine readable and AI-ready. However, as Albert Meroño-Peñuela and colleagues at the ODI note: "Modern AI is trained with datasets, and it has been shown that AI models are only as good as the datasets they are trained with. This means the NDL must not just be ordinarily data-ready, but…be AI data-ready" (10). Interoperability, metadata and dataset documentation, benchmarks and dashboards are key elements to consider to achieve this goal.

It is important that the NDL adopts open data standards and should be capable of providing live dashboards and AI model sandboxes for continuous evaluation. Meroño-Peñuela and colleagues propose "that the NDL does not limit itself to the publication of datasets, but that by acting as a sandbox it can also host a limited number of AI models pre-trained with NDL data that can be constantly evaluated" according to metrics, benchmarks and dashboards (10). Continuous monitoring and sandboxing will help protect data privacy and give the public assurances about the use of AI and their data. As well as providing AI models for research, AI can be used to improve data interoperability in the NDL. "AI can be used as a tool within the NDL to perform basic, routine data stewardship, maintenance, preparation, and quality tasks towards improving its interoperability and integration", write Meroño-Peñuela and co-authors (10).

## Financial sustainability

The technical design for the NDL will depend on the funding allocated to the initiative. This requires careful consideration by government. The NDL's costs include set-up costs for the

initial configuration of the technology and ongoing operational, governance and management costs. The government should develop flexible, long-term funding for the NDL to ensure its effectiveness and longevity. Experts noted that significant cost saving could be made if a federated approach to the NDL is adopted. Utilising existing services for areas including output checking, manual data preparation, and indexing would reduce the need for a large initial outlay.

However, a commercial approach might be needed for the NDL to be sustainable, which could involve charging a fee for data access based on factors like data volume and usage. "This approach would incentivise data providers to maintain high-quality, user-friendly data formats and allow users to choose the most suitable data extraction methods. A commercial model can drive significant benefits for the public, research and private sectors by streamlining data access and fostering innovation", write Browne and co-authors (9). Data can also be provided for a fee paid at the time of data sharing based on the data's assigned value and any egress cost. "The provider responsible for delivering the data egress should cover this cost, as well as any additional commercial fee. This approach allows the data user to choose how to access the data and ensures that the commercial entity reimburses the Government annually for the use of the data. Consequently, the Government would not need to enter into multiple contracts", note Browne and colleagues (9). However, while a charging model allowing the industry to access the NDL would be beneficial for the UK economy, opening access to the private sector would require strong safeguards.

## Summary

**Vision:** a clear vision is needed of how the NDL will serve research and public service delivery

**The Public**: the NDL must engage with and earn the trust of the public to safely deliver on its mission

**Federation:** to ensure robustness and operational resilience, the NDL should be built of individual federated elements in a stepwise manner

**Existing Infrastructure:** the NDL should build on data infrastructure already in place

**Governance:** aligning and standardising governance across the elements contributing to the NDL is essential to build trust and deliver operational efficiency

**Growth:** automated operations will be essential to the delivery and growth of the NDL

**Artificial Intelligence:** the NDL will be essential to the delivery of the UK's AI vision. AI should also be adopted to maximise the success of the NDL

**Sustainability:** the NDL cannot rely on everlasting public financing – it must engage with industry and safely deliver a self-sustaining business model

## Conclusion

The NDL has the potential to be a game-changing infrastructure for UK science, innovation and public services. Successfully established, it would be a unique resource for scientists and industry globally, given the UK's data-rich environment stemming from the NHS and other sources. Defining the precise scope and purpose of the NDL will be an essential first step in its development and will be crucial to informing the infrastructure's design. The public should be consulted and included in driving the vision for the NDL. As such, an urgent need exists for a series of public consultations and engagement activities to clarify the vision for the NDL. A dedicated communications team for the NDL also needs to be established to promote public engagement, understanding, and transparency, thereby fostering trust in the project.

In terms of design, the NDL should not be a huge database held in a single organisation: the government should learn from, use, and incorporate existing infrastructure into the NDL wherever possible. While an incremental delivery mode might be less attractive politically, it is technically advantageous and will reduce risk. A diverse, distributed infrastructure will avoid vendor lock-in and reduce the chance of failure of this large and complex project. Federation is supported by experts as the best design solution for the NDL. Various designs for a federated architecture can be considered including a modular approach, use of existing TREs, and a data fabric approach. Establishing an expert group on a federated architecture for the NDL would help decide the most appropriate data model. To support UK science and innovation as well as the AI Opportunities Plan, the government should ensure the NDL is AI-ready. This goal requires improving interoperability and ensuring data is machine readable. An open call for external experts and organisations to collaborate will help in achieving this outcome by bringing in the necessary level of expertise.

In collaboration with the Treasury and other relevant departments, the government needs to develop a sustainable funding approach for

the NDL. The business model needs to insulate this important initiative from changes in public sector priorities. Charging in some form seems a necessary step for sustainability and should be carefully considered by policymakers.

An end-user and stakeholder community committed to the success of the NDL already exists. Collaborating with these subject matter experts and the wider public is essential throughout the development stages of the NDL and will ensure the initiative's sustainability. Although the NDL has the potential to be a transformative initiative for the UK's health and the economy, much groundwork needs to be done. Moving forward with planning in the coming months will be essential for the NDL's timely delivery.

Finally, we should note that the UK data landscape is developing rapidly – for example the National Health Data Research Service has recently been announced as a £600m investment from government and the Wellcome Trust. Clearly the NDL will need to adapt to and capitalise on these developments. We predict an exciting period for data in the UK.

# References


1. Expert reaction to The Sudlow Review – Uniting the UK's Health data. Science Media Centre. Nov 2024. https://www.sciencemediacentre.org/expert-reaction-to-the-sudlow-review-uniting-the-uks-health-data/#:~:text=Prof%20Sir%20Martin%20Landray%2C%20Professor,to%20improve%20healthcare%20for%20decades.

2. COVID treatment developed in the NHS saves a million lives. NHS England. March 2021. https://www.england.nhs.uk/2021/03/covid-treatment-developed-in-the-nhs-saves-a-million-lives/

3. A Federated Architecture for a National Data Library. Amugi, M et al. DARE UK. Jan 2025. https://zenodo.org/records/14672004

4. Scientific use cases for cross-domain sensitive data research in the UK. DARE UK (2024). Zenodo. https://zenodo.org/records/14025303

5. Delivering an effective National Data Library. Hardinges J, Icebreaker One, Starks G. Jan 2025. https://zenodo.org/records/14674066

6. Uniting the UK's Health Data: A Huge Opportunity for Society. Sudlow, C. Nov 2024. https://zenodo.org/records/13353747

7. PEDRI. https://dareuk.org.uk/community-groups-listing/the-public-engagement-in-data-research-initiative/

8. A National Data Library: the Modular Approach. Goldacre B, Bacon S, Stokes P. Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford. Jan 2025. https://zenodo.org/records/14671714

9. UK National Data Library: Distributed Architecture for Research. Browne W, Jones, P, Fleming J, Chang W, Petter J, Emrys Health, Francis Crick Institute, Snowflake. Jan 2025. https://zenodo.org/records/14673493

10. How an AI-ready National Data Library would help UK science. Meroño-Peñuela A, Massey J, Newman A, Simperl E, Open Data Institute, Kings College London. Jan 2025. https://zenodo.org/records/14672231

**Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, infectious disease, and climate and health.**

**Wellcome Trust, 215 Euston Road, London NW1 2BE, United Kingdom**
**T +44 (0)20 7611 8888, E contact@wellcome.org, wellcome.org**